AD_____

Award Number: DAMD17-00-1-0707

TITLE: Automated Free Text Analysis Methodologies for the Total Army Injury and Health Outcomes Database

PRINCIPAL INVESTIGATOR: David Jensen, D.Sc.

CONTRACTING ORGANIZATION: University of Massachusetts
Amherst, Massachusetts 01003-3285

REPORT DATE: March 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2001	Final (1 Sep 00 - 28 Feb 01)

4. TITLE AND SUBTITLE
Automated Free Text Analysis Methodologies for the Total Army Injury and Health Outcomes Database

5. FUNDING NUMBERS
DAMD17-00-1-0707

6. AUTHOR(S):
David Jensen, D.Sc

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts
Amherst, Massachusetts 01003-3285

E-Mail: jensen@cs.umass.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

none provided

14. SUBJECT TERMS:
injury, health, outcomes database

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

This project investigated how data mining techniques could be applied to the Total Army Injury and Health Outcomes Database (TAIHOD), developed by the U.S. Army Research Institute for Environmental Medicine (USARIEM). In particular, we explored the potential for "information mining", a unique fusion of techniques from data mining with methods for automatically extracting data from the unstructured text reports that often accompany medical records. Data extracted from text may provide useful elements that could contribute to the predictive accuracy of statistical models constructed with data mining techniques.

Body

The project was largely a "proof-of-concept" for a new approach to data mining in medical data. The goals were to: 1) identify an appropriate data analysis task; 2) prepare data for that task; and 3) evaluate the performance of several data mining algorithms on those data. An appropriate task would be one that includes both structured data (e.g., a patient record with age, gender, and variables indicating medical condition) and unstructured text (e.g., a written narrative from a doctor). Our conjecture was that, by applying advanced information extraction algorithms to the unstructured text, additional variables could be created that would assist data mining. This would allow the construction of more accurate statistical models than could be built with the original structured data alone.

Unfortunately, the results of this pilot study were disappointing, largely because of: 1) extensive requirements for data anonymization, preparation, and cleaning; 2) difficulties hiring personnel to work on the project; and 3) ceiling effects on the selected task. Below we discuss the two early phases of the project (those that were most successful), briefly discuss the final phase (which was cut short by personnel departures and the ending of the project), and discuss possible future work.

Task Selection

The goals of the project required that we select a data mining task for which the data consisted of both highly structured records (the usual input to data mining algorithms) and unstructured text (not an input to traditional data mining algorithms). We explored a variety of alternative projects, including classifying the causes of drowning deaths and understanding the causes of parachute accidents. In addition, we considered the task of identifying gender based on first name and other demographic data. While this latter task did not involve large bodies of unstructured text, it would involve text (the names) and structured data (the demographic data).

We selected the task of understanding the causes of parachute accidents for three reasons. First, a similar data set had been previously studied by USARIEM personnel and had produced publishable results (Amoroso, Bell & Jones 1997). This published work would provide a basis for comparison and leverage existing expertise. Second, accurate statistical models could be extremely useful, given the potential to reduce injuries by targeting specific populations of trainees for extra training or supervision. Third, the data

set had both structured data and narrative descriptions of the accidents provided by instructors.

Data Preparation

This section briefly describes three steps used to prepare the parachute data for analysis. Such preparation is a large and important part of the process of data mining (Brodley & Smyth 1997; Fayyad, Piatetsky-Shapiro, & Smyth 1996). Specifically, we applied a three-step process to prepare the parachute data for analysis: 1) variable selection; 2) text anonymization; and 3) text variable extraction. The latter two processes involved developing novel software or using software previously developed at the University of Massachusetts' Center for Intelligent Information Retrieval (CIIR).

Variable selection

The data originally extracted for analysis contained 575 cases and 170 variables. The initial decision to include a variable in the extracted data set was based on the meaning of the variable, not on its intrinsic suitability for analysis.

We analyzed the distribution of each variable independently, to determine its suitability for analysis. In this process, we took into account the data mining algorithms we intended to apply later (e.g., algorithms for constructing decision trees and association rules) and the data requirements of those algorithms.

Of the original 170 variables, 112 had values for fewer than 100 cases. Of those, the vast majority had values in fewer than 40 cases. For example, values for the variable AGE was present in only 3 cases and the availability and use of goggles was known in only 3 cases. Such variables rarely provide substantial predictive power, and the algorithms we intend to use are not highly tolerant of missing values. All of these variables were eliminated.

Of the remaining variables, 18 had essentially only one value. Without substantial variation in the values of the variable, there is little point in including it in the analysis. These variables were eliminated.

After eliminating these two classes of variables, 40 variables remain, including one that is an ID variable. Four of the remaining variables appeared essentially useless for our purposes of analysis. The variables denoted the sequence number of accident, the file's source, the date, and whether the case was a member of duplicate set. These variables were eliminated. In addition, the variables TIME and HOUR were essentially identical, so HOUR was eliminated.

Six of the remaining variables weren't really useful for prediction. Instead, they were useful for constructing a "dependent" variable (what a statistical model is attempting to predict). The six variables are:

Name	Annotation
DAYHOSP	Days hospitalized
DAYLOST	Days lost
DAYREST	Days restricted
INJCOST	Total injury cost
NBPART1	Body part 1 affected
PINJCOST	Injury cost this person
Primary_Error01	Parachutist error 1
Primary_Error02	Parachutist error 2

These variables were eliminated from the set of predictors (but were used later to produce alternative dependent variables).

Two of the remaining variables encoded complex data which required recoding. Specifically, GRADE and MOS1T4. USARIEM staff provided procedures to recode the variables into discrete categories that are more amenable to analysis.

Text Anonymization

In addition to numeric and symbolic variables (e.g., age and rank, respectively), the parachute data contains text fields with narrative descriptions of the accident. One of the goals of the project was to use data derived from the text fields as part of the analysis. However, much of the textual data contains personal names and social security numbers (SSNs) of military personnel. Such personal identifiers are not necessary for the analysis, but they make it impossible for outside researchers (e.g., UMass) to work on the data without creating legitimate privacy concerns.

To alleviate such concerns, we wrote a program to automatically anonymize personal identifiers (names and SSNs) by converting them to surrogate names and numbers. The program accepts a text string, identifies each proper name and SSN, creates a substitute name or SSN for each, carries out the substitution, and then stores the anonymized text string. This substitution process preserves repeated uses of the same name (called "coreference" in the field of natural language processing), so the text remains meaningful to human readers.

This process was made more difficult because the text strings in the parachute data were stored in all capital letters, rather than upper- and lower-case letters which would have provided capitalization cues for proper names. However, the process was made easier by the rank identifications that often precede the first use of a proper name (e.g., "PFC Smith"). As a side benefit, this process converted the text to upper- and lower-case, to aid human readers.

The program went through five iterations where changes were made to the software at UMass, UMass personnel tested the program on ten text strings that had been hand-anonymized by USARIEM personnel, and then the program was tested on the full data at USARIEM. This process allowed the software to be developed at UMass without compromising the privacy of military medical records.

Text Variable Extraction

A large number of medical databases, including the parachute data, contain text fields that provide narrative descriptions of accidents or medical conditions. However,

essentially all data mining techniques assume that cases are represented by numeric and symbolic variables. Thus, one of the first steps in using text fields is to recode key aspects of the meaning of a narrative description into one or more numeric or symbolic variables.

The meaning of narrative descriptions are generally easy for humans to understand, but human language has proven surprisingly resistant to automated processing, despite several decades of work by computer scientists and linguists. That said, several relatively simple approaches based on statistical techniques have proven successful for limited tasks. In addition, "tagging" techniques that identify the linguistic structure of sentences has also been developed and successfully applied.

Our approach to extracting variables from text uses both parsing and some limited statistical approaches. Specifically, we use JTAG, a tagger that predicts the part-of-speech for each word in the sentence. JTAG was developed at UMass and has been used in a variety of systems for information retrieval and information organization.

After words have been associated with a particular part of speech, then useful variables are created based on frequently occurring nouns, verbs, and noun phrases. Specifically, if a word or phrase occurs in more than 10% of the records and less than 90%, it becomes the basis for creating a variable. Each variable created from the text fields are Boolean. That is, they have a value of 1 if the word or phrase occurs in that case's text field and they have a value of 0 if the word or phrase does not occur in the text field.

Analysis

Unfortunately, relatively little analysis was conducted on the parachute data set. In part, this was due to the extensive requirements for data anonymization, preparation, and cleaning, which took up the vast majority of faculty, staff, and student time devoted to the project. Due to the sensitive nature of the data, the anonymization program had to be developed at UMass, but tested at USARIEM, resulting in delays and occasional miscommunications. In addition, the task of data preparation was slowed by the extensive domain expertise required to understand the meaning of variables and their values. For example, merely understanding the GRADE variable required several days of work with frequent communication between UMass and USARIEM.

These difficulties were exacerbated by staffing problems. The project began at a time when many students were leaving or foregoing graduate school to pursue "dot com" jobs, and both CIIR and the Knowledge Discovery Laboratory (KDL) were experiencing difficulties hiring staff and graduate research assistants. In addition, due to several successful funding proposals, KDL was experiencing rapid growth in the demands on its students and staff.

That said, some initial analysis of the parachute data revealed some interesting associations. For example, jumps in Georgia were far more likely to be recorded with [error 6] (104/136=0.76) than jumps in North Carolina (126/287=0.44). The difference in error based on state is highly significant ($p \leq 0.0001$). Similarly, the type of error recorded for a given jump (Primary error 1) does not differ by military branch ($p = 0.6017$), but does differ slightly by grade ($p = 0.0328$).

As mentioned earlier, additional analysis of the data were cut short by the departure of key personnel (the principal research assistant assigned to the project) and the end of the project. In addition, some preliminary analysis indicated that it was unlikely that a highly

predictive statistical model could be constructed. This phenomenon often called a *ceiling effect*, referring to the existence of a theoretical upper bound on the performance of any statistical model.

Key Research Accomplishments

- Explored several potential data mining projects, ultimately settling on an investigation into the causes of injuries sustained during parachute training exercises. This task was selected after examining many alternatives, including correcting gender information in patient records based on name and demographic information, classifying the causes of drowning deaths, and understanding causal factors of automobile accidents. The primary reason for the selection was the availability of both structured data records and an accompanying unstructured text report for each record.
- Demonstrated that several data sets exist at USARIEM that would benefit from approaches that combine methods in data mining with those from information extraction.
- Developed a program for data santization and transferred it to USARIEM. The program identified names and SSNs of individuals in free text data and substituted consistent identifiers for these individuals. Version 1.0 was delivered in October 2000. Additional versions (version 1.1 – version 1.5) were delivered in October, November, and December. Such santization was a necessary first step for data to be used outside of USARIEM.
- Developed a data set for analyzing the causes of parachute injuries.

Reportable Outcomes

Due to the inconclusive results obtained during this pilot study, no publications resulted from the work on the parachute data set. The PI (Jensen) is a coauthor of a paper describing the application of classification tree learning to explore predictors of disability and hospitalization among active duty soldiers ([cite]). The principal authors of that paper are at USARIEM, though the PI has provided technical guidance on evaluating the results of applying the SAS Enterprise Miner software.

The contract supported the work of one graduate student, Sudheer Gaddam.

In addition, the work has led to a longer-term collaboration between USARIEM and one of the laboratories involved in the original contract, the Knowledge Discovery Laboratory (KDL). During the past six months, USARIEM and KDL personnel have begun planning a project to analyze the effect of the Gulf War deployment on the subsequent length of service of Army personnel. This project would make use of KDL's PROXIMITY software, which provides the ability to build statistical models from complex, relational data structures. The software escapes the assumptions of independent, identically distributed (i.i.d.) observations that is common to many statistical techniques. Analyzing such relational data may be essential to understanding the effect of deployment on length of service. We expect this work to generalize to a large number of other health-related analysis tasks.

Conclusions

The approach of combining data mining approaches with information extraction has promise. Regardless of the techniques used, the process of cleaning, processing, and preparing data can be time consuming and difficult, particularly when different project participants have expertise about the data and expertise about the analysis techniques, as was the case with this project. A longer-term relationship between KDL and USARIEM should allow additional transfer of expertise about data mining to USARIEM, development of several case studies, and use of some of the unique data mining algorithms developed at KDL.

References

Amoroso, P., N. Bell, and B. Jones (1997). Injury among female and male Army parachutists. *Avaiation, Space, and Environmental Medicine* 68(11):1006-1011.

Brodley, C. and Smyth, P. (1997). Applying classification algorithms in practice. *Statistics and Computing* 7:45-56.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). From data mining to knowledge discovery in databases. *AI Magazine*. Fall. 37-54.

Strowman, S., P. Amoroso, L. Senier, and D. Jensen (2003). The use of data mining methods to explore predictors of disability and hospitalization among active duty soldiers. Manuscript in preparation.

Appendices

Personnel

David Jensen
Sudheer Gaddam
Matthew Cornell